Sarcasm Detection & Sentimental analysis of Tweets Using Support Vector Machine.

[1]Shreyas Devarkar, [2]Karan Darji, [3]Ronit Bhatt, [4]Hezal Lopes
[1]B.E Student, [2]B.E Student, [3]B.E Student, [4]Asst. Professor.
[1]Dept. of Computer Engineering,
[1]UCOE, Mumbai, India

Abstract : - The growth of social media has been exceptional within the recent years. Vast quantity of knowledge is being place out on to the general public domain through social media. Twitter may be amongst the social giants that is a medium for showcasing information to the general public. Sentimental Analysis is that the best thanks to decide people's opinion concerning a selected post. The planned work presents analysis for sentimental behavior and wittiness detection in Twitter dataset. The planned work utilizes the support vector machine (SVM) to classify information into sarcastic and non-sarcastic and conjointly to classify the dataset into positive, negative or neural behavior. However, our major objective is to extend accuracy of finding of sarcasm or irony detection in tweets that is troublesome to detect as there's no face expression or intonation concerned in written information that eventually hampers the performance of sentiment analysis because the meant context of the text is scan in an exceedingly totally different manner by the machine in sarcastic or ironical texts.

***IndexTerms* - : Sentiment analysis , sarcastic , non-sarcastic , machine learning , social media , twitter , SVM , classification.**

I. INTRODUCTION

The contemporary world may be labeled as a digital world. With the invention of mobile devices and engineering, there has been an enormous pile of knowledge being generated by one device within the network. The web presently has over three billion connected devices across the world. With Associate in Nursing ever-increasing variety of devices, the quantity of knowledge being made is large. In modern-day humans are a unit pioneers in communication, developed countries boast concerning over hour of their population owning devices connected to the web. There has been a major rise in social media and microblogging sites like Twitter and Facebook within the last decade. These social media websites have provided associate in nursing open platform of communication within the era, various social media giants have additionally claimed to possess over one billion on-line users on one day. Twitter is claimed to cross over five hundred million tweets per day. The quantity of knowledge from such social media sites cause

a noteworthy chance, it unveils a variant and irregular dataset with form of information that is accessible publically.

Despite the supply of software package to extract information relating to a person's sentiment on a particular product or service, organizations still face problems relating to the information extraction. With the rapid climb of the planet Wide internet, individual's area unit victimization social media like Twitter that generates massive volumes of opinion texts within the type of tweets that is accessible for the sentiment analysis, this permits an enormous volume of data from an individual's viewpoint that makes it troublesome to extract sentences, read them, analyze them tweet by tweet, summarize them and organize them into a comprehensible format during a timely manner.

There are various different challenges that exhibit by streaming social media information. Informal language refers to the employment of colloquialisms and slang in communication, using the conventions of oral communication like 'would not' and 'wouldn't'. Not all systems area unit ready to observe sentiment from use of informal language and this might produce a haul for the analysis and decision-making method. Emoticons area unit a representation of human facial expressions, that within the absence of visual communication and prosody serve to draw a receiver's attention to the tenor or temper of a sender's nominal verbal communication, rising and ever-changing its interpretation.

For example, ☺ indicates a contented state of mind and indicates a tragic state of mind. Systems presently in situation don't have adequate information to permit them to draw feelings out of the emoticons. As humans typically started exploitation emoticons to properly categorical what they can't place into words. The information accessible is most frequently, only a few characters, that makes most text classification algorithms inefficient; as multiple keywords cannot typically be derived from such information. Another challenge is exhibit by the composition of information itself. Recent net culture has given rise to varied slangs and short forms like "LOL" (Laughing Out Loud) and "TTYL" (Talk to You Later) etc., Short-form is wide used even with

short message service (SMS). The usage of short-form is going to be used a lot of times on Twitter therefore on facilitate to attenuate the characters used, this can be as a result of Twitter has place a limit on its characters to one hundred forty.

Sentiment associate degree lysis has clad as an exciting new trend in social media with an outsized quantity of sensible applications that vary from applications in business to government use. Sentiment is associate degree angle, thought, or judgment prompted by feeling. Sentiment analysis is additionally called opinion mining. Sentiment Analysis is employed to classify the reviews exploitation the sentiment of the words into positive or negative or neutral. The feelings may be of any kind i.e. positive, negative, or neutral sentiment, or a numeric rating score stating the intensity of the sentiment. The most task is to accurately calculate the score of the tweet information and show the feelings therein explicit tweet.

II. LITERATURE REVIEW

In the last decade there features a nice rise within the scope for analysis of human behavioral on net information victimization machine learning. In [1], the authors have first shown a quick procedure to carryout sentiment analysis method to classify extremely unstructured information of Twitter into positive or negative classes. Secondly, have mentioned numerous techniques to carryout sentiment analysis on Twitter information as well as knowledge-based technique and machine learning techniques. Additional in [2], here the authors have analyzed the performance of Support Vector Machine (SVM) for sentiment analysis. For performance analysis of SVM, we've got used 2 pre- classified datasets of tweets, 1st dataset consisted of tweets relating to self-driving cars and second dataset was regarding the apple merchandise. Wood hen tool is employed for performance analysis and comparison. Results are measured in terms of preciseness, recall and f- measure. In [3], The paper presents a survey of sentiment analysis and classification algorithms. This survey complete that sentiment classification remains associate open field for analysis. There's tons of scope for algorithms in it. SVM and Naïve Bayes are most well liked algorithms for sentiment classification. Sentiment analysis of tweets is extremely common. Datasets from sites like Amazon, IMDB, flipkart are wide used for sentiment analysis. Deeper analysis is needed just in case of social networking sites. In several cases, context thought is extremely vital. Therefore, have same that there's additional analysis needed during this field. In [4], in this paper, the simplest way of up the existent with detection algorithms by as well as higher pre-processing and text mining techniques like emoji and slang detection are presented. For classifying tweets as disrespectful and non-sarcastic there are numerous techniques used, several of that are briefed in section two. However, the paper takes up a classification formula and suggests numerous enhancements, that directly contribute to the development of accuracy. The project derived analytical views from a social media dataset i.e., twitter dataset and additionally filtered out or reverse analyzed disrespectful tweets to realize a comprehensive accuracy within the classification of the information that's bestowed. The model has been tested in time period and might capture live streaming tweets by filtering through hashtags and so perform immediate classification. In [5], the work identifies only one variety of with that's common in tweets: distinction between a positive sentiment and negative scenario. It's bestowed a bootstrapped learning technique to amass lists of positive sentiment phrases and negative activities and states, and show that these lists are accustomed acknowledge disrespectful tweets. This work has solely damaged the surface of potentialities for distinctive with arising from positive/ negative distinction.

III. PROPOSED SYSTEM

With large variety of tweets and knowledge returning in at a brisk pace, it's vital to pick the most effective doable technique for the information obtained. One amongst the most effective doable machine learning strategies out there for sentimental analysis is Support Vector Machine (SVM). The planned system has three major steps (1) Data Pre- processing (2) Training the algorithm (3) Sarcasm Detection (4) Classifier.

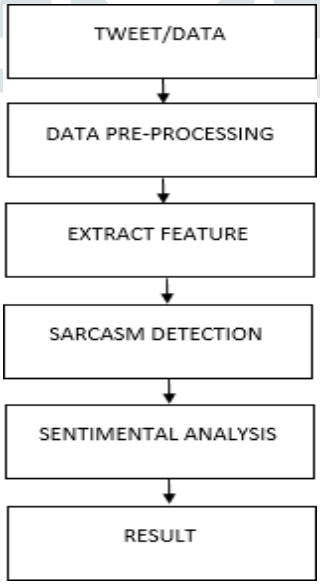

FIGURE III: SYSTEM ARCHITECTURE

(1) Data Pre-processing

Data pre-processing is that the most vital step within the sentiment analysis method. There square measure 3 steps during this stage, namely

a) Hashtag identification and replacement b) Slang dictionary mapping and c) Emoji dictionary mapping.

The dataset provided for the projected model could be a 2000 tweets dataset that contains general tweets with critical or non-sarcastic labels. This dataset could be a manually classified dataset that additionally is a brand new introduction during this paper. This assortment of tweets consists of the many things which require to be processed before moving onto the classification part. The primary step is hashtag identification and replacement. Hashtags square measure words or phrases preceded by a hash sign (#), used on social media websites and applications, especially Twitter, to spot messages a couple of specific topic. The hashtags within the dataset square measure lots. Some like "#elections2019" "#irony" etc., square measure helpful for classification however things like "#Lisbon" "#entertainment" aren't notably helpful. These hashtags square measure treated as traditional tokens and passed to a part of Speech tagging. The weight relies on the POS tag appointed. Successive step in pre-processing is that the emoji lexicon mapping. The favored trend of victimization emoji's in tweets can't be unheeded because it carries a great deal of weightage for classification. The emoji's in an exceedingly tweet square measure known so mapped with the manually designed emoji lexicon introduced during this paper. The lexicon contains the favored emoji's tagged as positive or negative or neutral.

The last step is that the slang lexicon mapping. The slang lexicon contains all the favored slangs and their meanings or full forms as key- value pairs. Once a tweet is being analyzed, if there's any slang that's detected it's like a shot mapped to the lexicon and replaced by the acceptable that means or full kind. This aids within the classification method.

(2) Data Preparation

After the pre-processing steps are completed, the info ought to be ready and did for the classification part. The tweets are a set of sentences that can't be directly fed into the classifiers. Hence, three major steps are performed to organize the info for successive part. The stages are (i) Word tokenization (ii) POS(Parts-Of-Speech) tagging (iii)Stemming and Lemmatization.

The first step is tokenization. Tokenization is performed on tweets to interrupt them down into excellent substantive modules from a sentence. Typically, tokens are often in terms of paragraphs or whole sentences except for the planned model it's a word. The tweet is lessened into words and also the keywords which is able to aid in classification are chosen and stop words are removed. when the tweets are tokenized, a part of Speech tagging is performed. The words during a tweet and their components of speech play a task in classification. If the person is employing a heap of adjectives there's an occasion that he's describing one thing with an excessive amount of praise, that hints concerning it being saturnine. With this in mind, the propose model tags the components of speech for every word. Stemming is constructed upon the thought that words with identical stem are progress that means, the words are stemmed to spot the words that are similar in that means. Lemmatization is the method of identification of the foundation word of the assorted words utilized in the tweet. As an example, words like mice are reborn to mouse. Such conversion clarifies the context of usage for the word and makes it easier to map it with its that means.

(3) SVM Classifier.

There square measure numerous SVM algorithms that uses completely different kernels supported the sort of knowledge to be classified. The Support Vector Machine are often viewed as a kernel machine. As a result, you'll modify its behavior by employing a completely different kernel operate. The foremost standard kernel functions are: 1. The linear kernel, 2. the polynomial kernel, 3. the RBF (Gaussian) kernel. The linear kernel is usually counseled for text classification. Most of text classification issues square measure linearly severable. The linear kernel is sweet once there's tons of options, that is as a result of mapping the information to a better dimensional house doesn't very improve the performance. In text classification, each the numbers of instances (document) and options (words) square measure massive. coaching a SVM with a linear kernel is quicker than with another kernel. significantly once employing a dedicated library. Support vector machine (SVM) solves the standard text categorization drawback effectively; typically outperforming Naïve Bayes because it supports the thought of most margin. The most principle of SVMs is to see a linear extractor that separates completely different categories within the search house with most distance i.e. with most margin. If we have a tendency to represent the tweet exploitation t, the hyper plane exploitation h, and categories employing a set into that the tweet should be classified, the answer is written as follows reminiscent of the sentiment of the tweet.

$$= \sum \mathrm{I}\,\mathrm{AI}\,\big(\mathrm{I}\,\vec{t}\,, \qquad \mathrm{AI} \geq 0$$

The idea of SVM is to work out a boundary or boundaries that separate distinct clusters or teams of information. SVM performs this task constructing a collection of points and separating those points exploitation mathematical formulas. Figure. 3.1 illustrates the info flow of SVM. Below fig. shows SVM progress

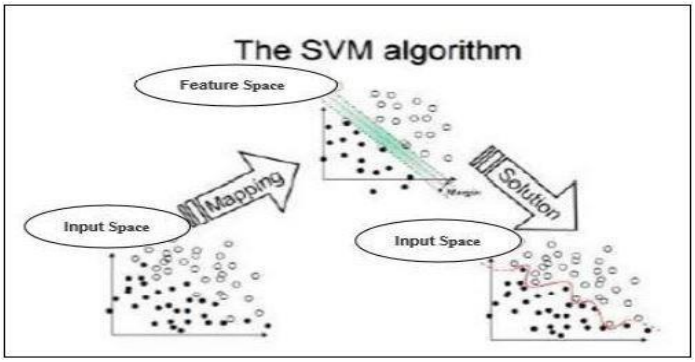

Figure. 3.1

After the SVM is applied the info is assessed into Positive Negative or Neutral reckoning on the score of the information.

(4) Sarcasm Detection

Sarcasm detection is one in all the foremost difficult tasks in machine learning. This is really the employment of remarks that clearly mean the other of what they assert, created so as to harm someone's feelings or to criticize one thing in a very risible approach. The use of irony is to mock or to mention one thing in an opposite approach of the particular intent. Sarcastic comments are tough to capture by humans and therefore may be a challenge for machine to accurately notice. The steps concerned in with or irony detection, which is as follows,

(A)Feature extraction

The terribly initiative in feature extraction is engineering the feature extraction. Here the unwanted or reedy information from is debarred and solely helpful data is maintained back. This feature extracted information is extracted in such the way that meaning of sentence remains preserved. Feature extraction uses tools like POS tagging, removal of characters, tokenization etc. That is analogous there to utilized in sentiment analysis. The feature values calculated, i.e. the feature extraction score is then mapped into vectors. These vectors are any saved into a wordbook known as the wordbook vector that is employed by the SVM classifiers. The information set (tweets) are divided into coaching and take a look on acting data and any test vector and train vector are calculated.

(B)Operations information Set (Training & Testing)

The data set is that the most vital part of the classification. The dataset used is of 10,000 tweets. Set is split into coaching and testing data set. The coaching of the information set starts with pre-assigning dataset with artificial weights. Victimization Linear kernel from SVM the feature score accuracy is calculated by cacophonic feature extracted dataset, fitting in SVM model. Accuracy score is calculated five consecutive times victimizing completely different split points.

(C)Validation

The trained SVM classifier is currently applied on the Testing knowledge and therefore the accuracy is calculated by cross validation and additional confusion matrix is planned. A confusion matrix could be a table that's usually describe the performance of a classification model (or "classifier") on a group which take a look at knowledge that verity values are identified. The confusion matrix itself is comparatively easy to know, however the connected language may be confusing.

	precision	recall	f1-score	support
Ironic	0.73	0.84	0.78	416
Non-Ironic	0.37	0.23	0.28	168
micro avg	0.67	0.67	0.67	584
macro avg	0.55	0.53	0.53	584
weighted avg	0.63	0.67	0.64	584

66.6095890410959

Figure 4.C.1

	precision	recall	f1-score	support
Ironic	0.74	0.70	0.72	412
Non-Ironic	0.36	0.41	0.38	172
micro avg	0.61	0.61	0.61	584
macro avg	0.55	0.55	0.55	584
weighted avg	0.63	0.61	0.62	584

61.130136986301366

Figure 4.C.2

Above is that the calculation of the performance of the trained SVM classifier. The F1-score could be a live of a test's accuracy. It considers each the exactitude p and therefore recall r, the take a look at to work out the score: p is the variety of correct positive

results divided by the amount of all positive results came by the classifier, and r is that the variety of correct positive results divided by the amount of all relevant samples. Within the planned work additional in confusion matrix to increase its performance, we've used and compared Normalized confusion matrix and confusion matrix.

In Figure. 4.1, The normalized confusion matrix, the rows are a part of the particular category, whereas the columns represent a part of the expected category. The diagonals represent a part of the expected category and the values of correct foretold category. The off-diagonal components represent the incorrectly foretold category that were erroneously foretold as another category. The properly foretold ironic knowledge tweets had 0.84 accuracies and 0.16 were incorrectly foretold.

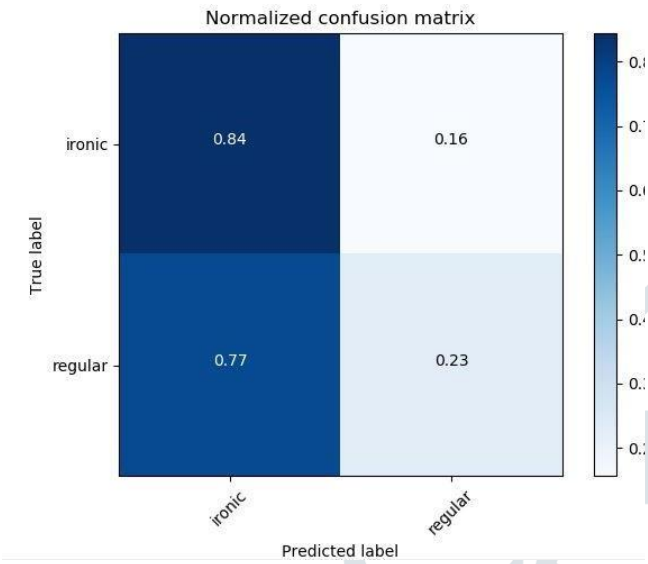

Figure 4.1

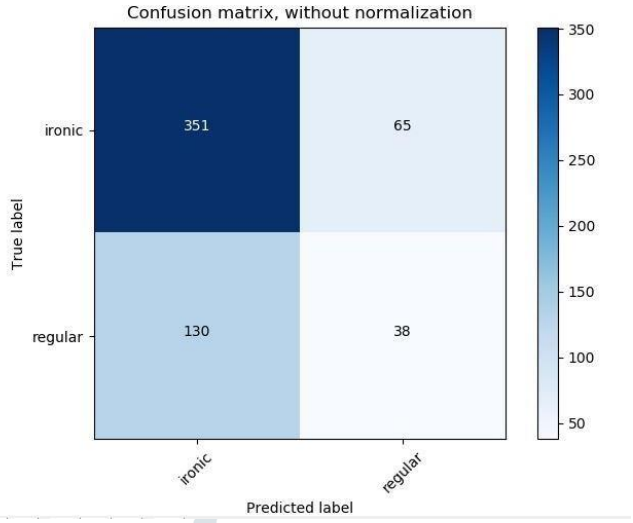

Figure 4.2

IV. Result and Discussion

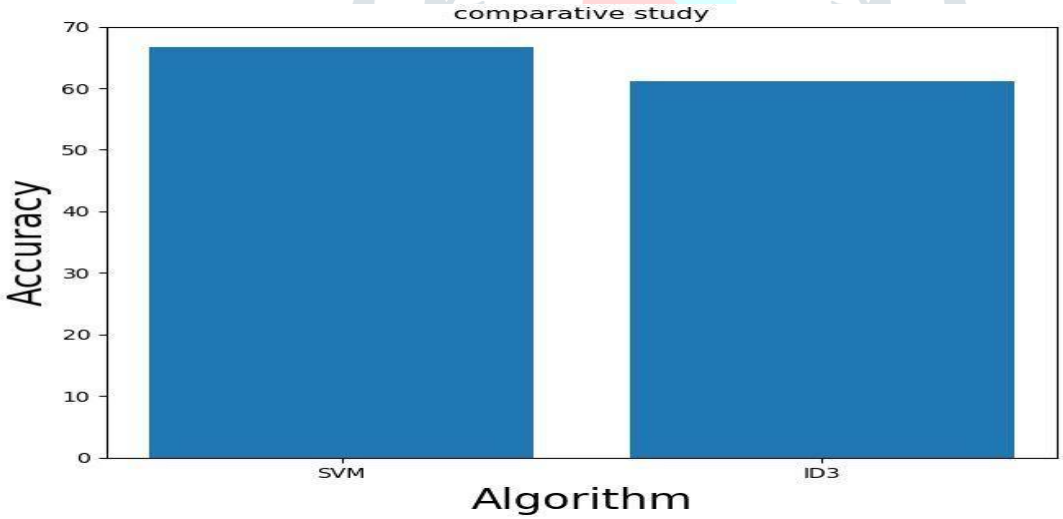

Figure. 5.1

(Figure. 5.1) The ID3 algorithmic program (i.e. call tree algorithm) was 61.13% correct whereas SVM and a much better and correct result with 69.06%.

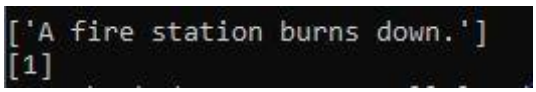

Figure. 5.2

In Figure 5.2 the SVM classifier properly detected the user driven knowledge as Ironic ([1]). From the results it's obvious that the SVM classifier performs comprehensively higher that alternative classifiers.

V. Conclusion And Future Scope

In this system we've given some way in machine learning technique, SVM may be applied to massive sets of knowledge to determine membership, during this case sarcastic(ironic) and non-sarcastic(non- ironic). The SVM classifiers perform rather well on the information used here. The SVM classifier will be an awfully smart job at learning and adapting the coaching information and

testing information. Comparison has been applied with the choice tree formula to match the accuracy between SVM and ID3(decision tree) formula. SVM classifier with success classifies positive, negative and neutral sentiment within the information tweets. However, information may be analyzed with the machine learning technique SVM taking into thought the slang, emoji employed in tweets lately.

In future there's have to be compelled to bring home the bacon higher accuracy. No Machine learning technique is 100% correct. here the planned system tries to realize most accuracy. There's heaps of future scope for sentimental analysis, thanks to ever increasing quantity of knowledge flowing in on the web. Hope to stay making an attempt and march towards 100% accuracy in predicting human sentiments.

VI. REFRENCES

[1] *Techniques for Sentiment Analysis of Twitter Data: A Comprehensive Survey* - Mitali Desai, Mayuri Mehta – issue of International Conference on Computing, Communication and Automation (ICCCA2016).

[2] *Sentiment Analysis of Tweets using SVM* - Munir Ahmad, Iftikhar Ali , Shabib Aftab – issue of International Journal of Computer Applications · November 2017

[3] *A Survey of Sentiment Analysis techniques* -Harpreet Kaur, Veenu Mangat, Nidhi – issue of (I-SMAC 2017)*Sentiment Analysis for Sarcasm Detection on Streaming Short Text Data*-Anukarsh G Prasad; Sanjana S, Skanda M Bhat, B S Harish – issue of 2017 2nd International Conference on Knowledge Engineering and Applications

[4] *Sarcasm as Contrast between a Positive Sentiment and Negative Situation*(2013) - Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva ,Nathan Gilbert, Ruihong Huang.

[5] *Sentimental Analysis Using Fuzzy and Naive Bayes*- Ruchi Mehra, Mandeep Kaur Bedi , Gagandeep Singh, Raman Arora, Tannu Bala, Sunny Saxena-issue of IEEE 2017 International Conference on Computing Methodologies and Communication.

[6] https://en.wikipedia.org/wiki/Global_Internet_usage#Internet_users

[7] www.internetlivestats.com/twitter-statistics/